

Research papers

A reliable linear method for modeling lake level fluctuations

Isa Ebtehaj^a, Hossein Bonakdari^{a,*}, Bahram Gharabaghi^b^a Department of Civil Engineering, Razi University, Kermanshah, Iran^b School of Engineering, University of Guelph, Guelph, Ontario N1G 2W1, Canada

ARTICLE INFO

This manuscript was handled by Marco Borga, Editor-in-Chief, with the assistance of Marco Toffolon, Associate Editor

Keywords:

Spectral analysis

Stochastic

SARIMA

Pre-processing

Lake level time series

Linear modeling

ABSTRACT

Accurate forecasting of lake level time series (LLTS) is an important but challenging problem with major economic, social and environmental implications. However, in recent years, the level of uncertainty in the existing LLTS forecast methods has increased significantly due to climate change, therefore, the need to develop more accurate models. The main research question for this study is whether it is necessary to use nonlinear methods in LLTS modeling or if linear methods can produce as accurate and reliable forecast tools. We introduce a new linear-based forecast method for LLTS using spectral analysis, seasonal standardization, and stochastic terms. The application of the new LLTS forecast method is tested on two case study Lakes, including the Van Lake, in Turkey and the Michigan-Huron Lake, in North America. A two-step preprocessing techniques based on standardization and differencing was used for the Van Lake, and spectral analysis and differencing was employed for the Michigan-Huron Lake. We then compared the accuracy and uncertainty of the proposed linear method with an artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS) methods. The uncertainty of the new linear LLTS forecast model was ± 0.00455 and ± 0.00264 for the Van Lake and the Michigan-Huron Lake, respectively, compared to ± 0.00625 and ± 0.00766 for the ANN and the ANFIS (respectively) at the Van Lake and ± 0.00312 and ± 0.00319 for the ANN and the ANFIS (respectively) at the Michigan-Huron Lake.

1. Introduction

According to the World Water Assessment Program (WWAP) report (UNESCO, 2003), by the year 2022, the amount of available freshwater across the globe will be reduced by 30–40%, a vital asset requiring superior administration for sustainable use. The inland freshwater lakes is one of the most critical assets found in the world. These lakes have a wide variety of beneficial uses and hence demands from recreation, fisheries, irrigation, drinking water, and hydroelectric energy production. However, to overcome the limitations of using the required complex input data sets and the limitations of the hydrological models, these phenomena are often modeled using historical data sets using a time series analysis and modeling techniques.

The existing methods in the prediction of the lake level time series (LLTS) are divided into two general categories: linear-based methods (i.e. stochastic models) such as auto-regressive integrated moving average (ARIMA), seasonal ARIMA (SARIMA) and so on, and the non-linear-based techniques emphasizing artificial intelligence (AI)-based methods. The use of linear methods requires a series of conditions, most notably the stationary nature of the series. So, for non-stationary time series such as most of the LLTS, it is necessary to employed methods for stationary the time series. Moreover, the existing of the different

deterministic term in LLTS which leads to the nonlinearity of the LLTS has led researchers to tend to use nonlinear methods.

There are many stochastic based studies in the field of hydrology (Valipour et al., 2013; Wang et al., 2015a; Traore et al., 2017; Myronidis et al., 2018; Kazemian-Kale-Kale et al., 2018; Zeynoddin et al., 2018), especially in lake level forecasting. Güldal and Tongal (2010) employed Fourier series to eliminate the periodic components of LLTS. The main weakness of their study is the not removing of trend term which is a deterministic term. Moreover, they did not survey the stationarization of time series using a quantitative test such as augmented Dicky-Fuller test. Existing such deterministic components led researchers to tend to the use of nonlinear methods in LLTS modeling.

Recently, the use of AI-based techniques as a power approach in solving complex nonlinear problems (Wang et al., 2009; Ebtehaj et al., 2015; Wang et al., 2015b; Khozani et al., 2017; Gholami et al., 2017; Atieh et al., 2017; Shaghaghi et al. 2017; Shabanlou, 2018; Freeman et al., 2018) remarkably increased and specially in lake level forecasting (Myronidis et al., 2012; Sanikhani et al., 2015; Shafaei and Kisi, 2016; Shiri et al., 2016; Zaji et al., 2018). Sanikhani et al. (2015) applied two different design of ANFIS, ANFIS with subtractive clustering (ANFIS-SC) and ANFIS with grid partitioning (ANFIS-GP), and gene expression programming (GEP) for lake level forecasting up to three

* Corresponding author.

E-mail addresses: bonakdari@yahoo.com (H. Bonakdari), bgharaba@uoguelph.ca (B. Gharabaghi).

month-ahead. They found optimal ANFIS-GP in predicting one and three months ahead and ANFIS-SC in predicting two month-ahead as the best models. However, the results of all model for predicting different month-ahead show the high level of accuracy in lake level predicting using these models.

A new method in the field of time series prediction is a combination of stochastic method and AI-based techniques. Moeeni et al. (2017a) by considering the different deterministic term of monthly inflow time series such as jump, trend and period, and removing these terms, proposed a novel hybrid method based on a combination of SARIMA model and ANFIS. The results approve the higher performance of the proposed method in comparison with individual ones, SARIMA and ANFIS. The presented hybrid methodology in Moeeni et al. (2017a) is performed with a combination of SARIMA and GEP and minor modification by Moeeni et al. (2017b). The results indicate the proposed hybrid methodology based on a combination of the linear and nonlinear method could be a high-performance method of prediction inflow time series.

Recently, a significant question was put forward in predicting soil temperature by Bonakdari et al. (2018). The authors in their paper claimed that the linear methodology could be a better approach to soil temperature modeling. The results of those paper proved their claim. The fundamental question raised is whether using the proposed methodology based on recognition and elimination of deterministic term of a time series and its linear modeling will lead to better results for other time series? This question is answered in the current study.

In this study, a novel hybrid linear-based methodology based on the stochastic models which are encoded in MATLAB software is presented for the first time for LLTS modeling. The proposed methodology is including of two-step preprocessing which is employed for the first time which is employed spectral analysis and standardization (seasonal and non-seasonal) techniques in combined with differencing (seasonal and non-seasonal) as the second step of preprocessing. Afterward, the remaining stochastic term of time series is employed as the input of stochastic-based models (i.e., ARIMA, SARIMA, etc.). The independence of the residual of the proposed linear-based methodology is examined by non-parametric Wald Wolfowitz run and cumulative periodogram tests. Moreover, the result of the linear-based methodology is compared with two published papers in high-quality journals, Çimen, and Kisi (2009) and Kişi (2009), and two popular artificial intelligence based techniques (ANN and ANFIS) using multi-criteria statistical indices.

2. Methods

This section includes two main sections; “proposed linear-based methodology” and “theoretical overview of nonlinear based techniques.” In the first section, the proposed methodology for LLTS is presented. At the beginning sub-section of Section 1, a different test is presented for the study of the time series length, and the existence of different terms of time series such as trend, jump and period. After identifying and removing any of these terms, the stochastic modeling is presented. Finally, the residuals independence related to all stochastic modeling is surveyed. In the second section, an overview of two important nonlinear based techniques (ANN and ANFIS) are presented, and the architecture of each model is given in the final sub-section.

2.1. Proposed linear-based methodology

2.1.1. Time series components

In hydrological studies, it is essential to include all periods of drought and wet in time series to ensure the credibility of studies. Therefore, Hurst's empirical relationship is used to determine the presence of boundary values in the time series. According to this relationship, if the coefficient is greater than 0.5, the length of the series is sufficient for modeling (Hurst et al., 1965).

Each time series has a special structure with its components of 4 different terms. These terms include trend, jump, period, and stochastic

term. Apart from the stochastic term that is required for modeling, the existence of other terms causes the time series non-stationary. Stationary of time series is the basic condition for stochastic modeling, which is examined using various tests such as Augmented Dicky-Fuller (ADF) and autocorrelation function. The ADF test relationships are as follows:

$$\Delta x(t) = \alpha + \beta_1 t + \beta_2 t^2 + \gamma l(t-1) + \varphi_1 \Delta l(t-1) + \dots + \varphi_{p-1} \Delta l(t-p+1) + \varepsilon(t) \quad (1)$$

$$\tau = \frac{\hat{\gamma}}{\sigma_{\hat{\gamma}}} \quad (2)$$

$$H_0: \gamma = 0 \text{ \& } H_1: \gamma < 0$$

This test is a single root test, which examines the stationary by probing the absence of a single root for the series. The $l(t)$ is the value of data at time t , τ is the statistic of the ADF test, $\hat{\gamma}$ is the estimated root coefficient, $\sigma_{\hat{\gamma}}$ is the standard error of estimated coefficient, H_0 is the assumption zero based on the existence of unit root and H_1 is the alternative assumption on the absence of single root, Δ is the first order differential operator, α is the constant value, β_1 and β_2 are the coefficient of the linear and the second degree (respectively) trend, which can be considered zero, φ_i is the i th auto-correlation coefficient, p is the maximum order for considered auto-regression and τ is smaller than the critical value, meaning the absence of a single root and stationary.

Since the zero assumption is based on the nonlinear series, the probability with the test statistic (P_{ADF}) is less than the significant level of 5% means that the time series is stationary. After reviewing the series, using the above test, in the case of non-stationary, it should identify the factors that have caused non-stationary. Therefore, Mann-Kendal (Hirsch and Slack, 1984), Mann-Whitney (Mann and Whitney, 1947), and Fisher (Kashyap and Rao, 1976) tests were used to examine the existence of each term of the trend, jump, and period (respectively) in the series.

2.1.2. Stochastic modeling

In stochastic modeling, deterministic components of time series should be identified and adequately eliminated, and only a stochastic term of the time series remains for modeling. In this case, the deterministic components are jumps, trends, and period. The above components are identified by the proposed tests and then eliminated by non-seasonal and seasonal standardization, spectral analysis, and differencing. The non-seasonal standardization relationship is presented as follows:

$$Std. = \frac{l(t) - \bar{l}}{S_d} \quad (3)$$

where Std is series converted with zero mean and standard deviation 1, $l(t)$ is the time series data, $\bar{l}$ the mean of data, and S_d is the standard deviation of the data. In the case of seasonal fluctuations, it can be applied as follows:

$$S.Std. = \frac{l(t, i) - \bar{l}(i)}{S_d(i)} \quad (4)$$

where $S.Std.$ is converted data with zero mean and standard deviation 1, $t(i, t)$ of data in t th year and i th season, $\bar{l}(t, i)$ mean of i th season data and $S_d(i)$ standard deviations of i th season data. Seasonal standardization leads to the elimination of correlations in the mean and standard deviations in a time series and stationary time series (Marco et al., 1993). In this method, it is no longer necessary to accurately identify the periodic components. Spectral analysis (S_f) is done that how much time series changes are affected by repetitive components. So, a spectrum is obtained over an interval of frequencies, and the periodic components can be distinguished in a non-stationary time series. With the help of spectral analysis based on the Fourier series expansion, a

time series can be decomposed into sinusoidal and cosine functions. In this way, the periodic component can be presented as a function of time, and by reducing the component of the time series, the periodic component is removed, and the series is stationary.

$$P(t) = \bar{l}(t) + \sum_{z=1}^k (\alpha_z \cos(2\pi f_z t) + \beta_z \sin(2\pi f_z t)) + \varepsilon(t), \quad t = 1, 2, 3, \dots, N \quad (5)$$

where $\bar{l}(t)$ is the mean of the original time series, $\varepsilon(t)$, the residual of the Fourier series extensions, α_z and β_z are Fourier coefficients, f_z is equal to z th harmonic of the base frequency, k is the maximum of the harmonic amount that $2z$ and $2z-1$ are for even and odd data respectively. The Fourier coefficients and f_z are calculated as follows:

$$\alpha_z = \frac{2}{N} \left(\sum_{t=1}^N l(t) \cos(2\pi f_z t) \right) \quad z = 1, 2, \dots, k \quad (6)$$

$$\beta_z = \frac{2}{N} \left(\sum_{t=1}^N l(t) \sin(2\pi f_z t) \right) \quad z = 1, 2, \dots, k \quad (7)$$

$$f_z = \frac{z}{N} \quad (8)$$

The periodic term is obtained by reducing the high quantity from the original time series as follows:

$$S_f = l(t) - P(t) \quad (9)$$

Moreover, the differential equation is as follows:

$$l'_t = l_t - l_{t-m} \quad (10)$$

where l_t is the amount of time series at time t and l_{t-m} is the amount of time series in m th delayed. Differentiation eliminates the trend and jumps in the mean and removes the periodicity of the mean. After deletion of definite components, using the SARIMA model coefficients, it is possible to establish a linear correlation between the time series values for predicting future values. The model's relationship is as follows:

$$\begin{aligned} & (1 - \phi_1 B - \phi_1 B^2 - \dots - \phi_p B^p)(1 - \Phi_1 B^S - \Phi_2 B^{2S} \\ & - \dots - \Phi_P B^{PS})(1 - B)^d(1 - B^S)^{D1}(f_t) \\ & = (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q)(1 - \Theta_1 B - \Theta_2 B^{2S} - \dots - \Theta_Q B^{QS})\varepsilon(t) \end{aligned} \quad (11)$$

where $l(t)$ is the amount of precipitation at time t , S is equal to periodicity, B is the differential operator, such that $B(x_t) = x_{t-1}$ and $(1 - B)^d$ is the d th non-seasonal differential, $(1 - B^S)^p$ is D th seasonal differential in step S , P and Q are the orders of the seasonal autoregressive and moving average parameters respectively, p and q are the orders of non-seasonal autoregressive and moving average parameters respectively, $\varepsilon(t)$ is the random component, Φ and Θ are the seasonal autoregressive and moving average parameters respectively, and ϕ and θ are the non-seasonal autoregressive and moving average parameters respectively.

2.1.3. Residuals independence

The independence of the residuals of the model is the correct modeling condition. Therefore, to check the independence of the residuals, non-parametric Wald Wolfowitz run test is used (Siegel and Castellan, 1988). This test is feasible for all hydrological events recorded in a series of times. To perform this test, first, to each data lower and higher than the mean value is given the characteristics s and b , respectively, then the statistics are calculated as follows:

$$U_R = (R - E_R)/S_d \quad (12)$$

$$E_R = (n_s + n_b + 2n_s n_b)/N \quad (13)$$

$$S_d = [2n_s n_b (n_s n_b - N)/(N^2(N - 1))]^{0.5} \quad (14)$$

where U_R is the test statistic, E_R and R are expected and achieved

sequences (respectively), S_d is the standard deviation of the sequences, N is the total number of data, n_s is the number of smaller data and n_b is the number of data is larger than the mean. If the probability corresponding to the test statistic (PUR) is greater than the significant level of 5%, the series of residues will be independent.

The flowchart of the proposed linear-based methodology in this study which is considered all of the above-mentioned statistical test and stationarization techniques, is presented as follows (Fig. 1):

2.2. Theoretical overview of nonlinear based techniques

2.2.1. Multi-layer perceptron neural network (MLP)

One of the most widely used and popular neural networks is multilayer perceptron (MLP) networks which include three main layers; input, hidden and output layers. In this study, the activation functions of “tansig” and “purelin” are considered for hidden and output nodes (respectively). During the MLP network training with backpropagation (BP), the computation is first performed from the input to the network output; then the calculated error values are propagated to the previous layers. Initially, the output is estimated in a layer-to-layer manner, and the output of each layer forms the input of the next layer.

The training steps with the help of this algorithm are: 1) Assign the weight matrix to each of the connections, 2) Select the input and output vector corresponding to it, 3) Calculate the output of the neuron in each layer, and thus calculate the output of the neurons in the output layer, 4) Update the weights by the method of propagating the network error (the difference between actual output and calculated output) to the previous layers, 5) evaluating the network performance trained using the mean square error (MSE), 6) returning to step (3) or the end of training.

2.2.2. Adaptive neuro-fuzzy inference systems (ANFIS)

The ANFIS method (Jang, 1993) is a combination of ANN and a fuzzy inference system (FIS). The first step to construct a fuzzy system was to determine the input parameters of the model and to select the type and number of membership functions. Additionally, to minimize the modeling error value, the number of significant iterations should be considered, so increasing this value does not have a substantial effect on the modeling results. The membership function shape is modified during the training process to obtain the optimal relationship between the input and output of the model, until acceptable convergence is achieved.

For a first-order Sugeno fuzzy model with two inputs, the if-then rules are defined as:

$$\text{Rule 1: if } x \text{ is } A_1 \text{ and } y \text{ is } B_1, \text{ Then } f_1 = p_1 x + q_1 y + r_1 \quad (15)$$

$$\text{Rule 2: if } x \text{ is } A_2 \text{ and } y \text{ is } B_2, \text{ Then } f_2 = p_2 x + q_2 y + r_2 \quad (16)$$

where A_i and B_i are linguistic labels such as “less” or “low” and p_1, p_2, q_1, q_2 is the parameter set which is determined through the training process.

The structure of ANFIS includes five different layers; fuzzification, product, normalization, de-fuzzification, and output (respectively). In the “fuzzification” layer, the membership degree of each is determined. The activation degree and firing strength related to each rule are calculated in the “product” and “normalization” layers, respectively. In the fourth layer, the de-fuzzification process is done. Finally, in the last layer, the summation of all input signals are determined in the “output” layer. The more detail of the ANFIS network is referred to Moradi et al. (2018) for more details related to the ANFIS model structure.

2.3. Uncertainty analysis for the proposed methodology predictions

In this sub-section, a quantitative evaluation of the uncertainties in the forecasting of the lake level is offered using the proposed linear based methodology for two lakes, Van and Michigan-Huron, versus two

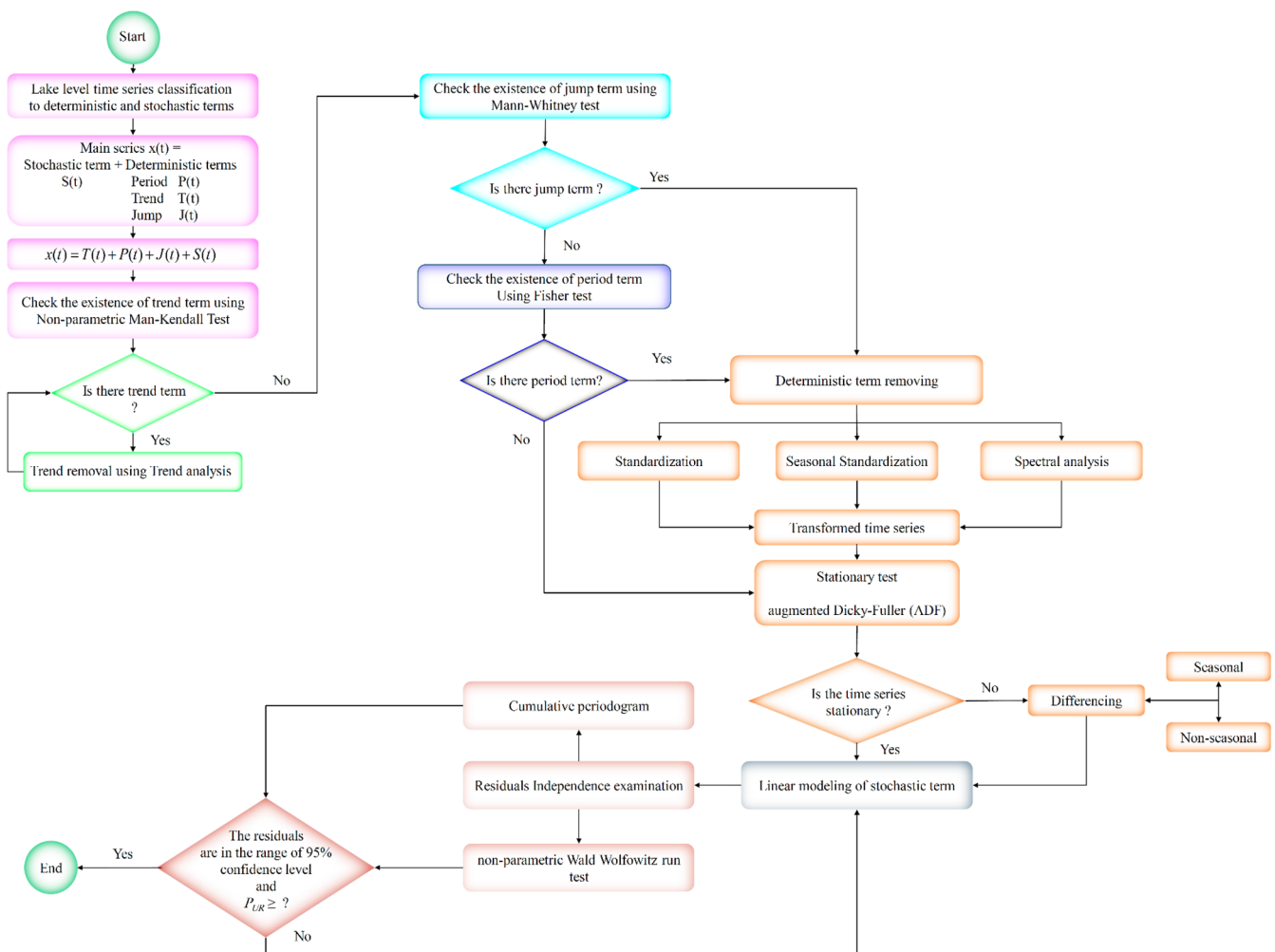


Fig. 1. The flowchart of the proposed linear-based methodology.

popular nonlinear based models, ANN and ANFIS. The provided results of uncertainty analysis (UA) in this subsection is applied to the test dataset (Ebtehaj et al., 2018; Azimi et al., 2018).

To survey UA, the individual forecasting error (IFE) is defined as follows:

$$IFE_i = F_i - O_i \quad (17)$$

where F and O are the forecasted and the observed samples, respectively. The IFE is employed for computing the mean of IFE ($MIFE$) and standard deviation of IFE ($SDIFE$). The $MIFE$ and $SDIFE$ are defined as follows:

$$MIFE = \sum_{i=1}^n IFE_i \quad (18)$$

$$SDIFE = \sqrt{\sum_{i=1}^n (IFE_i - MIFE)^2 / n - 1} \quad (19)$$

where n is the number of samples.

A positive (or negative) value of $MIFE$ shows the overestimate (or underestimate) the observed values. A confidence band around the forecasted values is defined based on the $MIFE$ and $SDIFE$ using the Wilson score method without continuity correction (WSMWCC). The 95% forecasting error interval (FEI) is achieved approximately by considering the ± 1.96 $SDIFE$.

2.4. Model verification

To evaluate the accuracy of the proposed linear methodology and the nonlinear methods (ANN and ANFIS), various indices are employed. The indices considered in this study include the coefficient of determination (R^2), two indices mean absolute percentage error ($MAPE$) and relative root mean square error ($RMSRE$) and the root mean square error ($RMSE$) index, which is an absolute index. The maximum value of R^2 is equal to 1, the higher the magnitude of this index for a model is closer to this number, indicating the high correlation of the predicted values using the model with observational values. The minimum values of the $MAPE$, $RMSRE$ and $RMSE$ indices are zero, and the maximum is infinite. The smaller value for these three indices, confirms the better accuracy of the model.

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left(\frac{|l_i - \hat{l}_i|}{l_i} \right) \quad (20)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (l_i - \hat{l}_i)^2} \quad (21)$$

$$RMSRE = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{l_i - \hat{l}_i}{l_i} \right)^2} \quad (22)$$

$$R^2 = \left(\frac{\sum_{i=1}^n (l_i - \bar{l}_i)(\hat{l}_i - \bar{\hat{l}})}{\sqrt{\sum_{i=1}^n (l_i - \bar{l}_i)^2} \sqrt{\sum_{i=1}^n (\hat{l}_i - \bar{\hat{l}})^2}} \right)^2 \quad (23)$$

where l_i the observed lake levels and $\hat{l}_i$ the standardized and differenced modeled lake level. It should be noted that the value of $RMSE$ has a dimension of the meter.

In addition to the indices, the Nash-Sutcliffe index is also used to evaluate the performance of the model for data that has a difference from the mean (Burnham and Anderson, 2002; Nash and Sutcliffe, 1970).

$$NSE = 1 - \left[\frac{\left(\sum_{i=1}^N (l_{0,i} - l_{f,i})^2 \right)}{\left(\sum_{i=1}^N (l_{0,i} - \bar{l}_{0,i})^2 \right)} \right] \quad (24)$$

where $l_{0,i}$ and $l_{f,i}$ are respectively the i th value of observed and forecasted lake levels. The Nash-Sutcliffe coefficient measures the forecasting power of models through the deviates from the mean value. The statistical value varies from $-\infty$ to 1 and the closer the coefficient is to 1, the more accurate model. Moreover, the Akaike's Information Criterion ($AICc$) is used to examine and compare the models by considering the accuracy and complexity of the model simultaneously. The $AICc$ allows comparing the statistical models by dealing with the compromise between the simplicity of the model and its goodness-of-fit. The $AICc$ index is defined as follows:

$$AICc = n \ln(\sigma_e^2) + \frac{2kn}{n - k - 1} \quad (25)$$

where K is the number of parameters, n number of months is the residuals' standard deviation.

3. Case studies areas

The data used in this study is the time series for the Van Lake in Turkey and Michigan-Huron Lake in North America. Van Lake is about 3713 km² in Turkey. The deepest part of the lake is 457 m, its longest dimension is 119 km, and the altitude is 1650 m. The water level reached its lowest level in the winter, and rising over the spring, feeding on melting snow, surrounding mountains, rainfall, and small rivers. The quality of water in this lake is not suitable for use in agriculture and drinking because of its high salinity. This lake is the largest lake in Turkey, and it has an essential role in its economy due to its proximity to Iran. During the last decade, the Van Lake level in eastern Turkey has been on the rise submersing the flat shorelines, damage to people's property, problems for the irrigation and drainage systems, and greatly concerning governmental officials and local administrators (Altunkaynak, 2007). Thus, the Van Lake level forecasting is a

necessary issue that attracts the researchers' attention.

The Michigan-Huron Lake is one of North America's largest water supply sources in both the United States of America and in Canada (Brinkmann, 2000). The surface area of this lake is 117,300 km² with a maximum depth of 281 m, the volume of water is 8460 km³. The largest inflow into this lake is St. Mary's River from Lake Superior and its main outlet is via the St. Clair River to Lake Erie. The population of the Michigan-Huron basin is about 33 million people which commercial activities are comprising production of industrial products, tourism and agriculture (Wilcox et al., 2007).

The position of the Van Lake and the Michigan-Huron Lake is presented in Fig. 2. The data related to Michigan-Huron Lake and Van Lake had been measured at Harbor Beach (United States) and Tatvan, (Turkey) stations, respectively. Van Lake data was measured from 1944 to 2002 (708 monthly levels). Michigan-Huron Lake water level data is also available from 1918 to 2013 (1152 monthly levels). This data has been measured at Hurbor beach station. For time series modeling, 80% of the beginning of the series, as the data (for the Van, 564 months and Michigan-Huron, 912 months), are considered for train and 20% end for testing (for Van, 144 months, and Michigan-Huron, 240 months). Measured data are hydrologically based, meaning from the beginning and the end months of October and September. Fig. 3 shows the time series of the data studied. In this figure, it can be seen that time series have trend and jump and severe fluctuations. However, for the further review, the autocorrelation function (ACF) and partial autocorrelation function (PACF) diagrams are presented in Fig. 4 illustrating the trends and fluctuations for the time series.

The statistical indices of the two lakes are presented in Table 1. In this table, from left to right, the number of data in both the test and train, the mean of the data ($\bar{x}$), the standard deviation (S_D), the coefficient of variation (C_V), the minimum (x_{min}), the first quartile (Q_1), the median (x_{50}), the third quartile (Q_3), the maximum (x_{max}), the skewness (S_k) and the kurtosis (k_u) of the both lakes data is seen. The mathematical forms of the mentioned statistics are defined as follows:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad (26)$$

$$S_D = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} \quad (27)$$

$$C_V = \frac{S_D}{\bar{x}} \times 100 \quad (28)$$

$$x_{50} = \begin{cases} x_{(N+1)/2} & \text{if } N \text{ odd} \\ x_{N/2} + x_{(N/2)+1} & \text{if } N \text{ even} \end{cases} \quad (29)$$

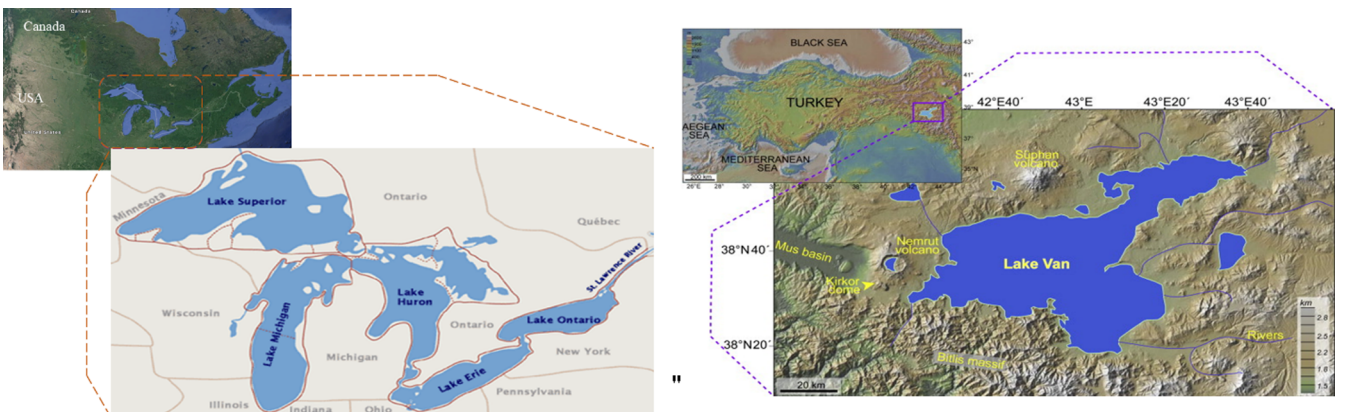


Fig. 2. Locations of Case Studies.

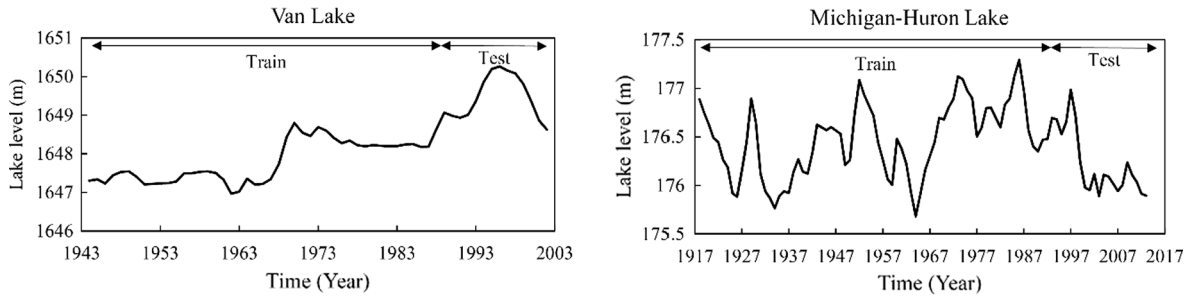


Fig. 3. Van Lake and Michigan-Huron Lake times series.

$$Q_1 \text{ and } Q_3 = x_{(y)} + z(x_{(y+1)} - x_{(y)}) \quad (30)$$

$$S_k = \frac{N}{(N-1)(N-2)} \sum_{i=1}^N \left(\frac{x_i - \bar{x}}{S_D} \right)^3 \quad (31)$$

$$k_u = \frac{N(N+1)}{(N-1)(N-2)(N-3)} \sum_{i=1}^N \left(\frac{x_i - \bar{x}}{S_D} \right)^4 - \frac{3(N-1)^2}{(N-2)(N-3)} \quad (32)$$

where N is the number of samples, x_i is the i th sample, y is the truncated integer value of w is equal to $L(N+1)/4$ where L is 1 and 3 for Q_1 and Q_3 (respectively), and z is the fraction component of w that was truncated away. It should be not that when w is an integer, $y = w$, $z = 0$, and Q_1 or $Q_3 = x_{(y)}$

It is seen that the values of the indices and the range of numbers for the two lakes are quite different. Distribution of data of both lakes with each other, in test and train intervals are different and asymmetric. The data kurtosis is negative in both lakes and is less than normal. The amount of data deviation from the mean is low, and this value is close to the test and train periods.

4. Results and discussion

4.1. The results of lake level time series modeling

The data of two lakes were evaluated by the “Hurst test” to verify the sufficient length of the data. According to the results, the coefficient obtained for the Van Lake is 1.07 and for the Michigan-Huron Lake is 0.85, both values are greater than 0.5. Therefore, the length of the statistical period is sufficient for modeling. Concerning Fig. 3, trends, jumps, and period in each case can be seen clearly. This is confirmed in Fig. 4 with the drawing of series ACF and PACF diagrams. In this figure, correlations of the same direction and periodicity represent the trend and period in the series. But for further investigation, the tests presented in the previous sections are applied to the series. According to the results of the tests presented in Table 2, the p-value of the Mann-Kendal, seasonal Mann-Kendal, Mann-Whitney tests for the Van Lake and the Michigan-Huron Lake time series are less than the significant level of 5% and confirm the trend and jump in the series. But surprisingly, Fisher's test data for each lake time series is less than 3, which means that there are no alternatives in the series. To the occurred contradiction, the spectrum of the series was drawn in Fig. 5.

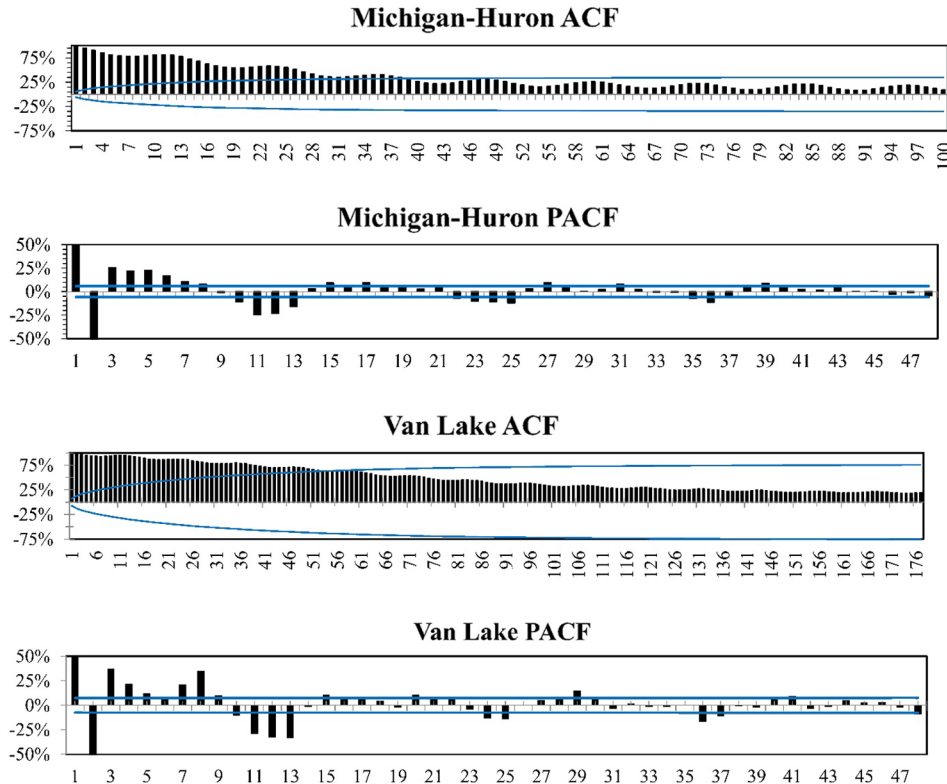


Fig. 4. The auto correlation function (ACF) and partial auto correlation function (PACF) of Van and Michigan-Huron lakes.

Table 1
Statistical indices of lake levels in training and testing phase.

Lake	phase	Num.	$\bar{x}$ (m)	S_D (m)	C_V	x_{min} (m)	Q_1 (m)	x_{50} (m)	Q_3 (m)	x_{max} (m)	S_k	k_u
Van	train	564	1647.85	0.61	0.37	1646.68	1647.34	1647.74	1648.33	1649.26	0.30	−1.00
	test	144	1649.54	0.59	0.35	1648.41	1648.99	1649.63	1650.08	1650.53	−0.16	−1.33
Michigan-Huron	train	912	176.47	0.40	0.16	175.58	176.16	176.48	176.76	177.50	−0.02	−0.69
	test	240	176.21	0.35	0.12	175.57	175.98	176.10	176.39	177.19	0.81	−0.04

Table 2
Test results of lakes before and after first step pre-processing.

Lake\Tests		Jump	Trend	Period		Stationarity
		%P _{MW}	%P _{MK}	%PS.MK	F [*]	%P _{ADF}
Van Lake		0.01	0.01	0.01	0.0916	86.87
1st step preprocessing	Std.	0.01	0.01	0.01	−7.19E+08	86.87
	S.Std.	0.01	0.01	0.01	−5.99E−12	86.87
	S _f	0.01	0.01	0.01	−4.17E−04	93.07
Michigan-Huron Lake		0.01	0.01	0.01	0.1046	15.91
1st step preprocessing	Std.	0.01	0.01	0.01	−6.40E+06	15.91
	S.Std.	0.01	0.01	0.01	−5.68E−12	2.09
	S _f	0.01	0.01	0.01	2.74E+02	7.92

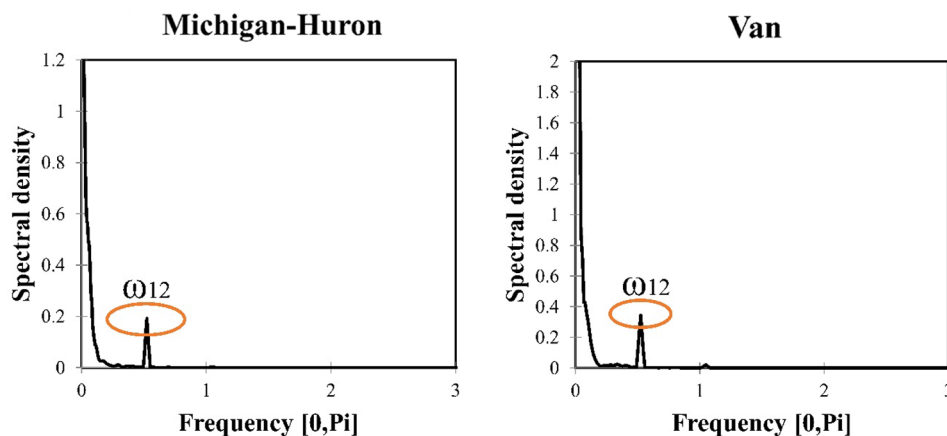


Fig. 5. Spectral density of lake levels.

In this figure, it can be seen that both series have a period at the frequency equivalent to step 12. On the other hand, there are apparent fluctuations in the graph of both time series. The Fisher test, therefore, has poorly tested the frequencies in series, and both series have a period in step 12. The probability corresponding the ADF test statistic to both series is higher 5%, indicating that both series are non-stationary. Therefore, in the structure of both time series, there are all definite components, trends (seasonal and non-seasonal), periodicity and jump. Consequently, they should be eliminated by appropriate methods to remain stochastic term in the series for modeling. For this purpose, in the first step, by a stationary using *Std.*, *S.Std.* and *S_f* endeavored to remove definitive terms. The test results are presented in Table 2 showing that none of the methods have been able to eliminate definite terms and have decreased even in *S_f*. In the meantime, the series from *S.Std.* for the Michigan-Huron Lake, according to the statistical probability of jump tests and trends is less than 5%, the stationary of the series has increased, and the probability of the test statistic has fallen below 5%, i.e., 2.09%, which is a contradiction. However, the results indicate that a pre-processing step is not enough for the time series and requires more pre-processing.

In the second stage of preprocessing, from the series of single stage preprocessing, non-seasonal and seasonal differencing, were made. The results are presented in Table 3. For Van Lake's time series, none of the

differencing methods on the time series of non-seasonal standardization have been able to eliminate jump in the series. Except for the seasonal differentiation in the seasonal series of seasonal standardization, the jump in the remaining Van Pre-processed series has been well removed, and the differencing methods on the spectral analysis series have the highest probability of statistics than standardizations. The trend has been entirely eliminated by using all three differential methods and in this case, the non-seasonal differencing for all three series from *Std.*, *S.Std.* and *S_f* has the best performance in eliminating the non-seasonal trend and the seasonal and non-seasonal differencing for the above series has had the best performance in eliminating the seasonal process. Fisher test, which in the previous step showed the series without a period, at this stage shows the series of non-seasonal differencing, seasonal differencing on *Std.* and *S.Std.* (respectively) and the non-seasonal and seasonal differencing simultaneously on the *S_f* in a periodic manner. However, all three series produced for the Van Lake are well stationary concerning the ADF test.

For the Michigan-Huron Lake, similar to the Van Lake, the three differencing methods have been applied to the series of non-seasonal standardization in the removal of jump performed weakly. The non-seasonal differencing series from *S.Std.* and the seasonal differential on the *S_f* series has also failed to remove the jump. But the method of non-seasonal and seasonal differencing simultaneously for *Std.* and *S.Std.*

Table 3
Test results of lakes related to second pre-processing.

		Lakes	Van			Michigan-Huron		
		Tests	Std.	S.Std.	S_f	Std.	S.Std.	S_f
2nd step preprocessing	Non-seasonal Differencing	% P_{MW}	0.01	19.08	85.3	0.01	0.01	23.74
		% P_{MK}	85.18	71.73	64.18	65.71	27.29	71.92
		% $PS.MK$	44.86	46.89	44.57	36.76	29.62	36.64
		F^*	752.2653	10.5807	416.1753	−0.0504	−0.0399	−0.0058
		% P_{ADF}	0.01	0.01	0.01	0.01	0.01	0.01
	Seasonal Differencing	% P_{MW}	0.01	0.01	94.02	0.01	55.49	0.18
		% P_{MK}	33.31	38.68	24.21	0.34	0.32	0.23
		% $PS.MK$	44.71	20.56	16.91	0.42	0.42	0.19
		F^*	108.5239	112.2988	−0.0079	−0.0002	−0.0001	−5E−05
		% P_{ADF}	0.01	0.01	0.01	0.01	0.01	0.01
	Non-seasonal & Seasonal Differencing	% P_{MW}	0.01	56.29	88.99	0.01	57.01	73.14
		% P_{MK}	57.97	53.98	37.91	92.39	95.55	66.52
		% $PS.MK$	61.24	61.5	72.41	93.12	80.53	85.99
		F^*	−0.0187	−0.0072	60.324	1.73E+03	1.81E+03	610.7359
		% P_{ADF}	0.01	0.01	0.01	0.01	0.01	0.01

removed this term well. The seasonal differencing method has not been able to eliminate the trend in the lake's pre-processed series. But the simultaneous differencing (simultaneous non-seasonal and seasonal) has eliminated this term well. The results of the Fisher test also show that the simultaneous differencing induces a periodicity in all three series from $Std.$, $S.Std.$ and S_f , but the test values for the other two methods of differencing are still less than 3. According to the results of the ADF test, the same as the Van, all three are stationary.

The ACF diagrams of the above series are plotted in Fig. 6. In this figure, it is seen that single-stage preprocessing is not capable of securing the series and there are many correlations in the series. By applying differentiation on the series, except for the non-seasonal differential in the Std series, and S_f , which caused severe seasonal correlations, the correlations value for other series declined sharply, and all have been stationary. On the other hand, the ADF test is not able to identify the seasonal correlations generated in the non-seasonal differential series on $Std.$ and S_f and they're non-stationary.

According to the results of the tests in Table 4 and the diagrams in Fig. 6, the series of simultaneous differencing methods are very suitable for modeling. In this method, it can be seen that not only have the acceptable results obtained in the tests but also, according to the diagrams, fewer parameters for modeling is needed. For example, in the seasonal differencing method, at least 11 non-seasonal parameters and 1 seasonal parameter for modeling are needed (for both lakes). While for simultaneous differencing, this number has reached a maximum of 2 parameters for the Van Lake and maximum 4 parameters for the Michigan-Huron Lake.

Considering the lower number of parameters for modeling that yields acceptable results leads to a simpler linear model that makes use of it more practical. Therefore, modeling is done according to the minimum values for the linear model parameters (Eq. (11)). Thus, for modeling of lake time series, the SARIMA model with seasonal and non-seasonal differencing and seasonal and non-seasonal parameters with a maximum order of 4, i.e., p, q, P and $Q = \{0, 1, 2, 3, 4\}$ are used (Fig. 6).

The results of the superior model indices for each of the methods are presented in Table 4. In this table, it is seen that the modeling has a very good coefficient of determination. The amount of errors is also very low. In this modeling, the combination of spectral analysis and then simultaneous differentiation has the highest correlation in the Van Lake and the Michigan-Huron Lake ($R^2 = 0.9954$ (Van); $R^2 = 0.9889$ (Michigan-Huron)) and the Std method for Van lake, and $S.Std$ method has the lowest value between all methods.

The Nash-Sutcliffe index values also show that the modeling power of the series is high, and the spectral analysis method with the SARIMA

model (3,1,1) (2,1,3)₁₂ for Van and the SARIMA model (1, 1,0) (0,1,1)₁₂ had the best performance for the Michigan-Huron Lake so that the value of this criterion is the least among the other two methods. While for the Van the best model, is the model (1,1,2) (0,1,1)₁₂ that in the first stage of its pre-processing the Std method has been used. The number of parameters given in this model is lower than the other models, and the values of the indices for this method are very close to the spectral analysis method. Since the values of the indices $R^2 = 0.9951$; $RMSE$ (m) = 0.0417; $NS_E = 0.9950$ for this model is very close to the spectral analysis method and the values of other error indices, $MAPE = 0.0019$ = $RMSRE = 0.0025$ is the same, so the model is chosen as the superior model. After identifying different models, the independence of the residuals of each model should be evaluated.

Generally, the best model for Van and Michigan-Huron Lakes is as (1,1,2)(0,1,1)₁₂ and (1,1,0)(0,1,1)₁₂, respectively, which demonstrate the seasonal and non-seasonal differencing is employed simultaneously in both lakes. The $AICc$ index which considers both the complexity and accuracy of each model for selecting the superior models for the Van Lake and for the Michigan-Huron Lake is −906.36 and −1580.11, respectively.

Moreover, the order of the non-seasonal autoregressive (AR) parameter for the Van Lake is one while the order of the seasonal AR is zero which indicates the seasonal AR is not used by the superior model for the Van Lake. The order of the non-seasonal and seasonal moving average (MA) in the proposed linear model for the Van Lake is 2 and 1, respectively. Indeed, the sum of the order of seasonal and non-seasonal AR and MA parameters related to a superior model for the Van Lake is four (AR = 1; MA = 2; seasonal AR = 0 and seasonal MA = 1). The order of the seasonal and non-seasonal AR parameter, as well as the seasonal MA for a superior model of the Michigan-Huron Lake is similar to the Van Lake. Moreover, this model did not employ the non-seasonal MA parameter for the Michigan-Huron Lake level prediction.

Consequently, the proposed linear method forecasted not only the fluctuations of the Van Lake and the Michigan-Huron Lake with a high level of accuracy but also the proposed model is straightforward with the lowest order of the seasonal and non-seasonal AR and MA parameters.

To investigate the independence of the remaining series and to ensure the elimination of periodicity detected in the series, the non-parametric Wald Wolfowitz run test (Table 4) was applied to the residuals series and the Cumulative periodogram drawn (Fig. 7). It is observed that the Wald Wolfowitz test values for all series of residues

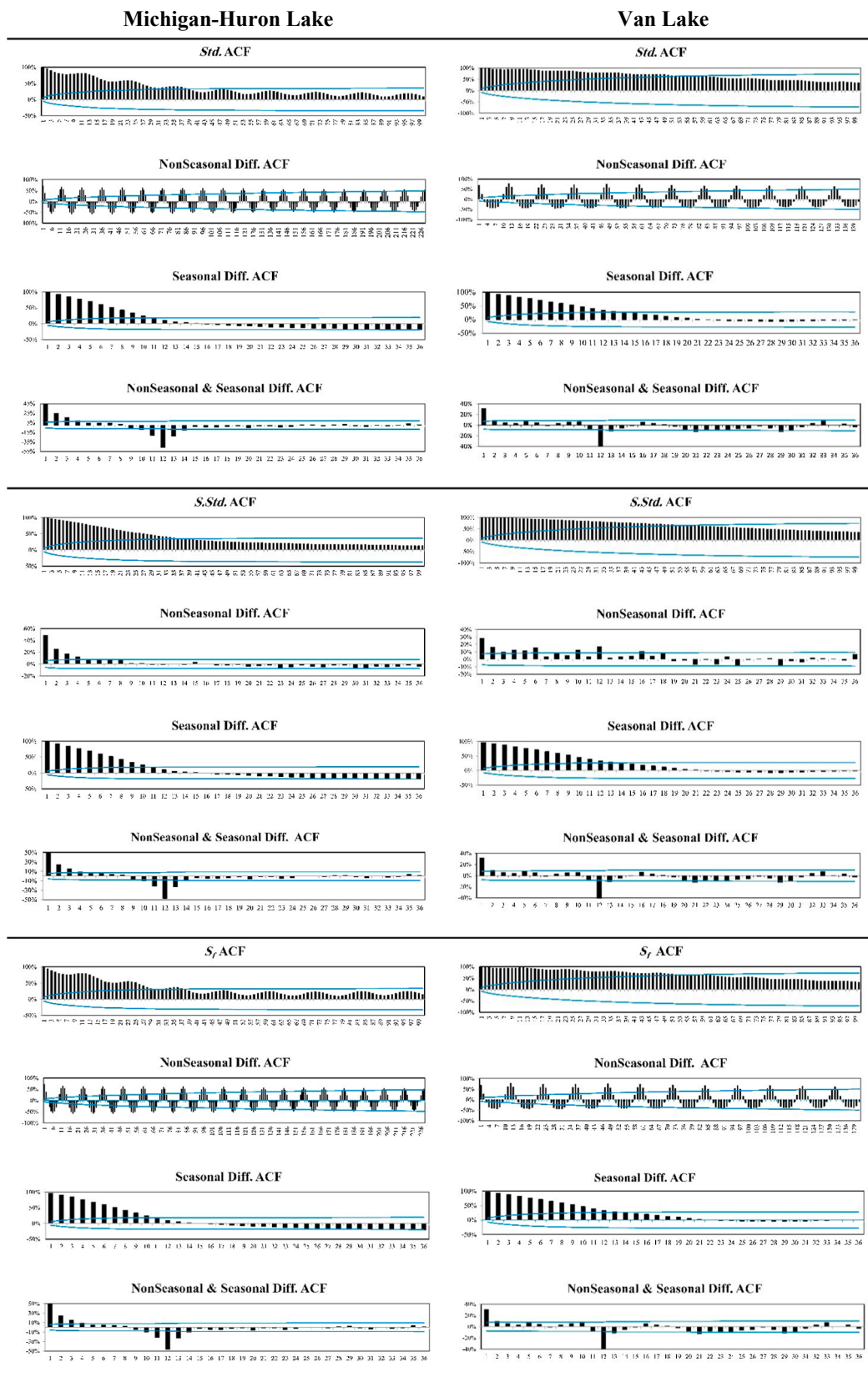


Fig. 6. Auto Correlation Function (ACF) of preprocessed data.

are more than 50%, indicating the relative independence of the residuals and the validity of the modeling. On the other hand, according to Fig. 7, all independent series and periodicity have been eliminated well.

The time series graphs for the Van Lake and Michigan-Huron Lake, along with the superior method for each lake, is presented in Fig. 8. In this figure, it can be seen that the modeling of the series results from the pre-processing using the *Std* and *S_f* method for the Van Lake and the

Table 4
Lake levels model results.

Lakes			MODEL	R^2	RMSE (m)	MAPE%	RMSRE	NS_E	AICc	%P _{UR}
Non-seasonal & Seasonal Differencing	Van	Std.	(1,1,2)(0,1,1) ₁₂	0.9951	0.0417	0.0019	0.000025	0.9950	−906.36	67.64
		S.Std.	(2,1,3)(0,1,2) ₁₂	0.9947	0.0433	0.0020	0.000026	0.9946	−889.25	95.02
		S_f	(3,1,1)(2,1,3) ₁₂	0.9954	0.0407	0.0019	0.000025	0.9952	−903.53	67.80
	Michigan-Huron	Std.	(3,1,4)(0,1,0) ₁₂	0.9779	0.0529	0.0242	0.000301	0.9765	−1395.03	50.86
		S.Std.	(1,1,0)(0,1,1) ₁₂	0.9888	0.0370	0.0172	0.000210	0.9885	−1577.06	53.02
		S_f	(1,1,0)(0,1,1) ₁₂	0.9889	0.0368	0.0171	0.000209	0.9887	−1580.11	74.51

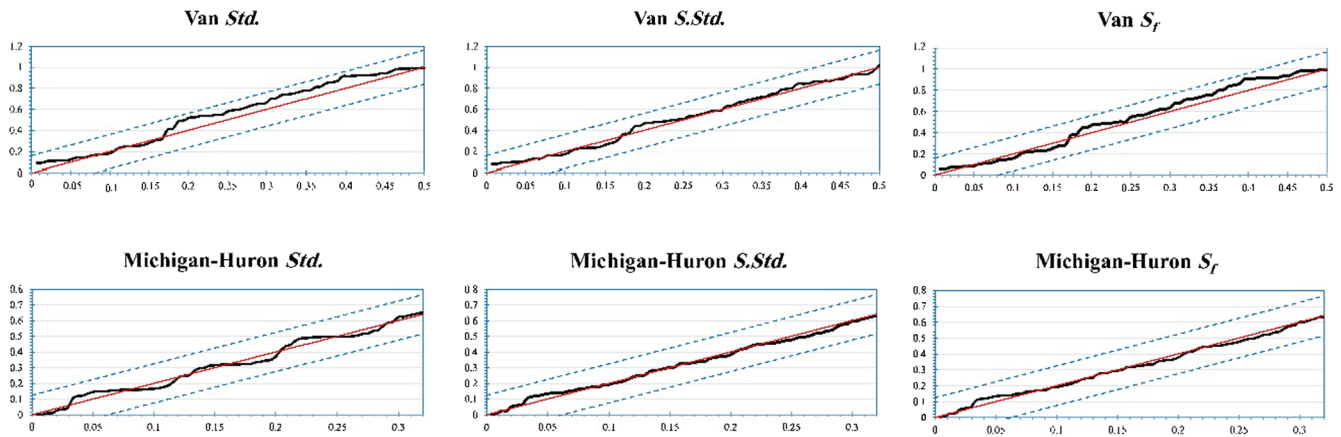


Fig. 7. Cumulative periodogram of lake levels models.

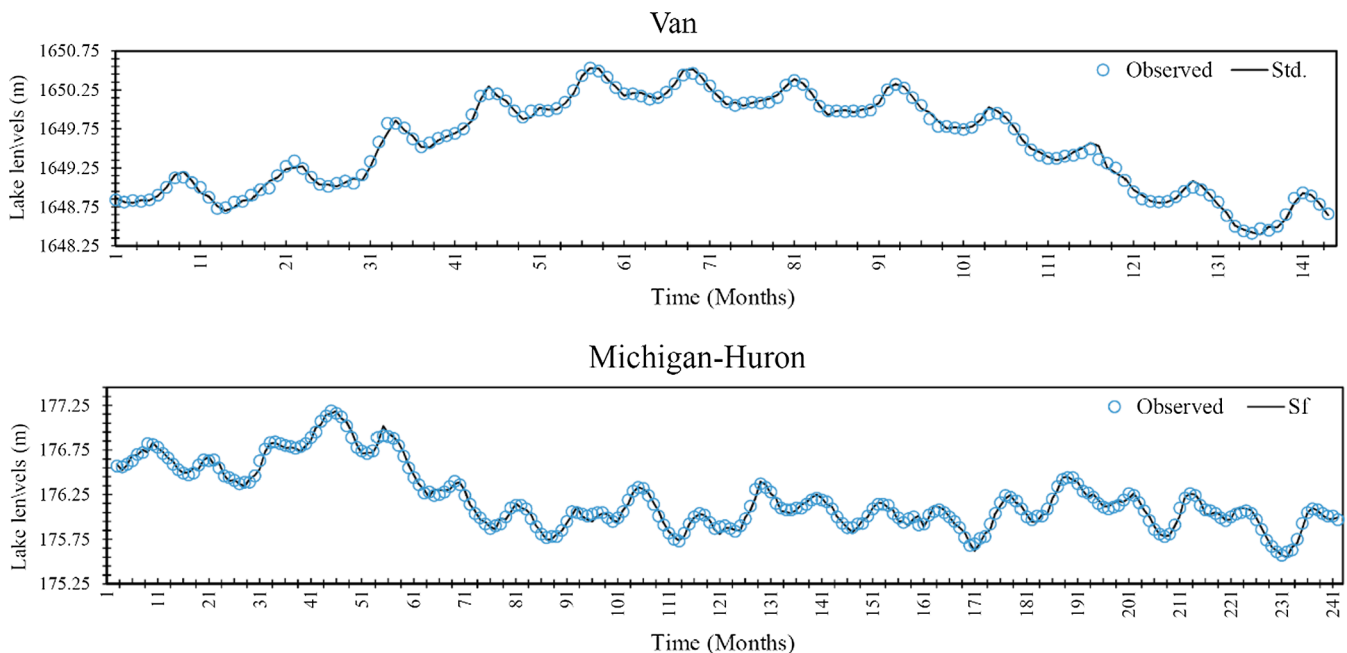


Fig. 8. Monthly lake levels estimate of superior preprocessing method in test period.

Michigan-Huron Lake, using the SARIMA model, provide very satisfactory results so that the predicted value has very good conformity with the observed values.

In Fig. 9, the box plot of the series is modeled using the SARIMA model for each of the 3 preprocessing methods. In this figure, it can be seen that the box areas for all three methods of stationary are very close to observational values, and the statistical characteristics are well maintained and close to real values. On the other hand, the predicted peak samples in Michigan-Huron Lake by Std. and S_f is much closer to the observed values related to S.Std.

4.2. Comparison of the proposed methodology with ANN and ANFIS

It is essential to adjust the modifiable parameters related to each model to achieve an accurate and optimal nonlinear based models (ANN & ANFIS). The optimal values of the ANN and ANFIS for modeling the water level of the Michigan-Huron Lake and the Van Lake which are attained through a trial and error process have been presented in the following table (Table 5):

Table 6 compares the results of the best models presented using the proposed linear based methodology for Van and Michigan-Huron lakes

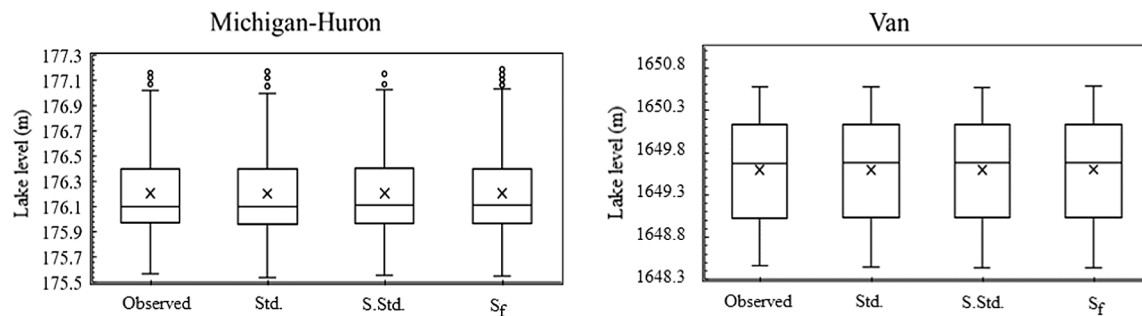


Fig. 9. Box plot of observed data vs. modeled data in test period.

Table 5

The optimal values of adjustable parameters related to different AI-based techniques.

Method	Parameter	Setting	
		Michigan-Huron	Van
ANN	Activation function	tansig	tansig
	Number of hidden layer	1	1
	Number of hidden neurons	1	1
	Training algorithm	BP	BP
	Iteration number	2000	2000
ANFIS	MF number	2	2
	FIS generation technique	FCM	FCM
	Training algorithm	Hybrid (BP-LS)	Hybrid (BP-LS)
	Iteration number	5000	5000

with the results of nonlinear methods (ANN & ANFIS). The results of linear methods are divided into two categories, results presented in previous studies (Kisi, 2009; Çimen and Kisi, 2009) and the results of modeling in this study. The Van Lake level modeling, presented by Kisi (2009) and Çimen and Kisi (2009), has a lower accuracy than the modeling performed in this study. Çimen and Kisi (2009) have been able to increase the accuracy of the Van Lake level modeling by increasing the number of hidden layer neurons and also considering a delay higher than Kisi (2009). The relative error presented for the ANN model (4,8,1) (Çimen and Kisi, 2009) is about one-third the value of this index for ANN (3,1,1) (Kisi, 2009). On the other hand, the coefficient of determination also increased by about 3%.

In this study, the non-linear modeling of these data has been modeled due to the lack of access to the exact amounts of modeling data in past studies to compute indices such as NS_E , AIC_C , $RMSRE$. The indices presented for Van's Lake data (test period) shows that the two models ANN (3,1,1) and ANN (4,1,1), which have a very simple structure, have

fairly similar accuracy, and the increase in lag does not have a significant effect on modeling. Also, for the data of this lake, two models with ANN models (similar regarding number and type of parameters) are modeled using ANFIS. The ANFIS model (FCM, 2,3) indicates that the fuzzy c-means clustering method has been used for FIS generation in ANFIS. The number of clusters considered is 2, and the number of inputs of the model is also 3 ($t-1$, $t-2$, $t-3$). For ANFIS models, with the constant of modeling conditions, the increase in the number of modeling parameters from 3 to 4 has led to a relatively small accuracy increase, although it also adds to the complexity of the model. Comparing the two models of ANFIS (FCM, 2,3) and ANFIS (FCM, 2,4) using the AIC_C index, which also considers the complexity of the model in addition to considering the accuracy of the model, shows that the AIC_C for ANFIS (FCM, 2,3) is less than the ANFIS (FCM, 2,4) ($AIC_C = -738.72$). The comparison of two nonlinear ANN and ANFIS methods using the AIC_C index shows that the ANN model (3,1,1) ($AIC_C = -797.86$) has the best performance among all the nonlinear models presented in this study. In this study, in addition to the statistics indices that only consider the model's accuracy ($R^2 = 0.9951$; $RMSE$ (m) = 0.0417; $MAPE$ % = 0.0019; $RMSRE$ = 0.000025; NS_E = 0.995) have better performance than nonlinear methods, in terms of $AIC_C = -906.36$, also have a significant advantage over nonlinear methods.

Therefore, the proposed linear-based methodology is a superior modeling technique compared to the existing non-linear methods by providing a simpler yet more accurate model than the nonlinear models. The results of the proposed linear-based methodology in the Michigan-Huron Lake and the Van Lake level, each of which in three different structures and are compared with two existing non-linear models in Table 6.

In both ANN and ANFIS methods, with the constant of all the parameters of the model, increasing an input parameter leads to an increase in modeling accuracy. It should be noted that the excessive increase in the number of inputs increases the complexity of the model. For this reason, the AIC_C index, which takes into account the

Table 6

Comparison of Proposed linear based methodology with two famous nonlinear methods (ANN & ANFIS).

Lake	Reference	Method	Inputs	R^2	$RMSE$ (m)	$MAPE$ %	$RMSRE$	NS_E	AIC_C
Van	Current study	Proposed methodology	(1,1,2)((0,1,1) ₁₂)	0.9951	0.0417	0.0019	0.000025	0.9950	-906.36
		ANN (3,1,1)	$t-1$, $t-2$, $t-3$	0.9898	0.0607	0.0029	0.000037	0.9891	-797.86
		ANN (4,1,1)	$t-1$, $t-2$, $t-3$, $t-4$	0.9896	0.0601	0.0029	0.000036	0.9896	-794.95
		ANFIS (FCM,2,3)	$t-1$, $t-2$, $t-3$	0.9893	0.0873	0.0044	0.000053	0.9762	-738.72
		ANFIS (FCM,2,4)	$t-1$, $t-2$, $t-3$, $t-4$	0.9895	0.0763	0.0037	0.000046	0.9821	-730.70
		ANN (3,1,1)	$t-1$, $t-2$, $t-3$	0.951	0.4021	0.0181	-	-	-
	Kisi (2009) Çimen and Kisi (2009)	ANN (4,8,1)	$t-1$, $t-2$, $t-3$, $t-4$	0.985	0.0199	0.0066	-	-	-
Michigan-Huron	Current study	Proposed methodology	(1,1,0)(0,1,1) ₁₂	0.9889	0.0368	0.0171	0.000209	0.9887	-1580.11
		ANN (3,1,1)	$t-1$, $t-2$, $t-3$	0.9837	0.0445	0.0201	0.000252	0.9834	-1485.01
		ANN (4,1,1)	$t-1$, $t-2$, $t-3$, $t-4$	0.9845	0.0433	0.0199	0.000246	0.9842	-1495.37
		ANN (5,1,1)	$t-1$, $t-2$, $t-3$, $t-4$, $t-5$	0.9855	0.0417	0.0192	0.000237	0.9853	-1508.74
		ANFIS(FCM,2,3)	$t-1$, $t-2$, $t-3$	0.9826	0.0459	0.0207	0.000260	0.9826	-1436.20
		ANFIS(FCM,2,4)	$t-1$, $t-2$, $t-3$, $t-4$	0.9840	0.0440	0.0201	0.000250	0.9838	-1443.41
		ANFIS(FCM,2,5)	$t-1$, $t-2$, $t-3$, $t-4$, $t-5$	0.9849	0.0427	0.0197	0.000243	0.9847	-1441.31

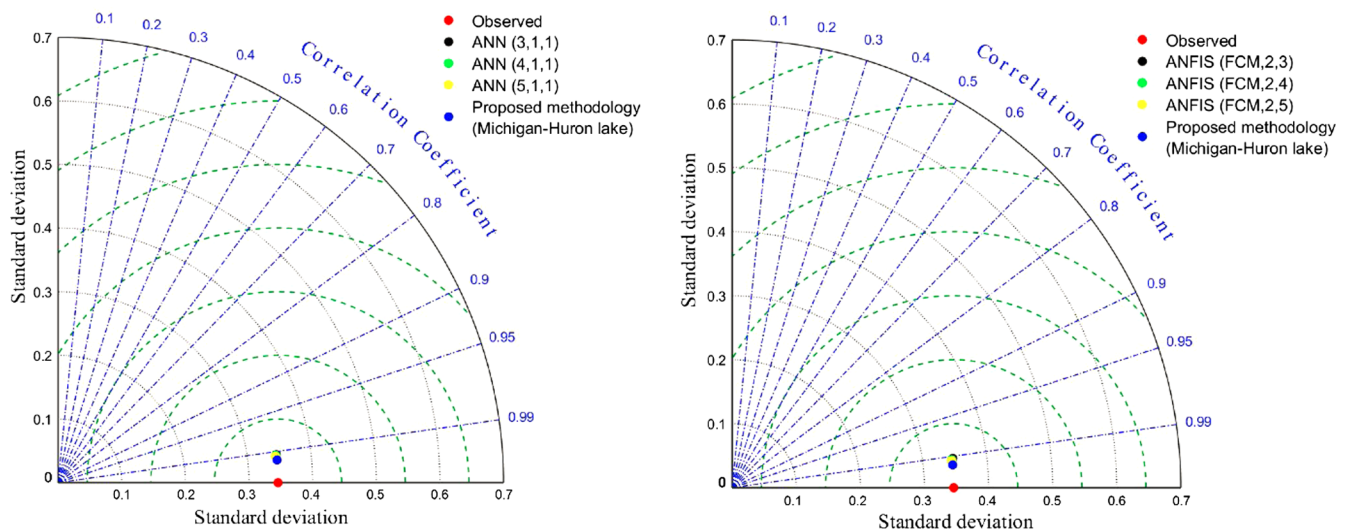


Fig. 10. Taylor diagram presentation for the proposed linear-based equation and nonlinear methods (ANN & ANFIS) at Van and Michigan-Huron lakes.

complexity and accuracy of the model simultaneously, is presented in this study. The lowest value of the AIC_C index is obtained in ANN and ANFIS methods for a different structure.

The use of the 5 input parameters leads to the best answer in ANN (5.1.1) ($AIC_C = -1508.74$). While using 4 parameters is along with the best performance in ANFIS (FCM, 2.4) ($AIC_C = -1443.41$). Similar to the Van Lake, the use of the ANN is better than ANFIS, so that the AIC_C index is less than the best ANFIS for the best ANN model, about 4%. However, other values of indices also generally outperform the ANN as compared to ANFIS. The comparison results of the proposed linear based methodology for Michigan-Huron lake ($R^2 = 0.9889$; $RMSE$ (m) = 0.0368; $MAPE\%$ = 0.0171; $RMSRE$ = 0.00209; $NS_E = 0.9887$; $AIC_C = -1580.11$) with nonlinear models shows the superiority of the proposed methodology over the existing methods.

The Taylor diagram (Taylor, 2001) is an appropriate tool for evaluating various methods. The values of the coefficient of determination in the form of the circular radius on its arc, the standard deviation values of two concentric circles relative to the center of the circle and the values of $RMSE$ in two concentric circles relative to the reference point of the hollow circle on the horizontal axis are plotted. The reference point indicates the position of the synoptic station based on the standard deviation of its time series because the $RMSE$ value and the coefficient of determination of the time series of the synoptic station will be zero and one compared to itself, respectively, so its position on the horizontal axis will be determined based on the standard deviation. The evaluation method in this diagram is that the position of the data under investigation based on $RMSE$, its coefficient of determination with the synoptic station of each data, and the standard deviation of the

time series on the diagram, which location is closer to the reference point on the graph, estimates the lake level value with more accuracy.

For the Van Lake level time series, comparing the results of the two models presented for each of the ANN and ANFIS methods with the proposed linear based methodology, it can be seen that the distance of the reference point from the proposed method to the non-linear method, relative to the observational data points is less likely to be due to better performance of this method than ANN and ANFIS.

For the Michigan-Huron Lake level time series, the points of the various models in the ANN and ANFIS methods show no significant difference, and sometimes these points are on each other, which indicates the insignificance of the difference in models. In fact, increasing the number of inputs will not have a significant impact on the modeling results. In this time series, like the Van Lake, the proposed linear based methodology to the ANN and ANFIS methods is closer to the reference point for observational data (Fig. 10).

The UA results have been presented in Table 7. This table represents the number of samples (NS), mean of individual forecasted error ($MIFE$), the standard deviation of individual forecasted error ($SDIFE$), the width of uncertainty band (WUB) and 95% forecasted error interval (FET). It is clear that the proposed linear based methodology technique for two lakes has performed better than nonlinear methods with less computed uncertainty.

The positive value of $MIFE$ for all linear and nonlinear models for both of lakes shown the overestimate performance of these models in lake level forecasting. The $MIFE$ for the proposed methodology is calculated as 0.03346 and 0.02922 compared to 0.04750 for ANN (4,1,1) and 0.03381 for ANN (5,1,1) for the Van Lake and the Michigan-Huron

Table 7
Uncertainty analysis for proposed linear based methodology versus nonlinear models.

Lake	Model	NS	MIFE (m)	SDIFE (m)	WUB	95% FEI
Van	ANN (3,1,1)	144	0.04747	0.03790	± 0.00625	(0.04123 0.05372)
	ANN (4,1,1)	144	0.04750	0.03700	± 0.00610	(0.04140 0.05359)
	ANFIS (FCM,2,3)	144	0.07191	0.04973	± 0.00819	(0.06372 0.08010)
	ANFIS (FCM,2,4)	144	0.06060	0.04647	± 0.00766	(0.05295 0.06826)
	Proposed methodology	144	0.03346	0.02757	± 0.00455	(0.02892 0.03801)
Michigan-Huron	ANN (3,1,1)	240	0.03535	0.02707	± 0.00344	(0.03191 0.03879)
	ANN (4,1,1)	240	0.03504	0.02548	± 0.00324	(0.03180 0.03828)
	ANN (5,1,1)	240	0.03381	0.02454	± 0.00312	(0.03069 0.03693)
	ANFIS (FCM,2,3)	240	0.03647	0.02788	± 0.00355	(0.03292 0.04001)
	ANFIS (FCM,2,4)	240	0.03550	0.02603	± 0.00331	(0.03219 0.03881)
	ANFIS (FCM,2,5)	240	0.03468	0.02504	± 0.00319	(0.03150 0.03787)
	Proposed methodology	240	0.02922	0.02076	± 0.00264	(0.02658 0.03186)

Lake (respectively). Moreover, the min value of WUB was calculated for the proposed methodology at the Van Lake ($WUB = \pm 0.00455$) and Michigan-Huron ($WUB = \pm 0.00264$). The smallest and the highest value of WUB are ANN (4,1,1) ($WUB = \pm 0.00610$) and ANFIS (FCM,2,4) ($WUB = \pm 0.00766$) for the Van Lake and ANFIS (FCM,2,5) ($WUB = \pm 0.02504$) and ANN (3,1,1) ($WUB = \pm 0.02707$), respectively. The proposed methodology had the lowest $MIFE$ and the smallest WUB of all the compared methods.

5. Conclusions

We presented a novel linear-based model for Lake Level Time Series (LLTS) forecasting and evaluated the performance of the methodology using two case studies of the Van Lake, in Turkey and the Michigan-Huron Lake, in North America. The LLTS were preprocessed using, standardization, seasonal standardization, spectral analysis, and seasonal and non-seasonal differencing. The new LLTS modeling technique presented in this study outperformed two popular nonlinear techniques, the ANN and the ANFIS, with improved accuracy and lower uncertainty. The main reason for the improved performance of the new model is due to the appropriate pre-processing by identification of the deterministic terms and removing them before stochastic modeling. The main finding of the study are as follows:

- The Fisher's test was poor in detecting the periodicity in the LLTS while using the spectral analysis of different series was a better method for identifying the seasonal periodicity.
- We employed the standardization, seasonal standardization, and the spectral analysis to remove the deterministic terms from the LLTS. However, none of these methods were able to remove the deterministic terms adequately. Therefore, the first round of preprocessing was not adequate and a second round was necessary to reach optimum results.

Appendix A. Hurst coefficient

The Hurst coefficient is defined as follows:

$$H_C = \log(R/S_x)/\log(N/2) \quad (A1)$$

$$S_x^+ = S_{\min}^{\max}(S_1, S_2, \dots, S_N) \quad (A2)$$

$$R = S^+ + |S^-| = (S_{\max} - S_{\min}) \quad (A3)$$

In the above relationship, S_x is the standard deviation of the data and N is the total number of data, $S_{\min}^{\max}$ is the maximum and the minimum of the cumulative deviate series as $S = \sum_{i=1}^N l_i - \bar{l}$.

Appendix B. Trend analysis

To test the changes in the hydrological time series, there are many statistical tests that one of the most popular and most widely used tests is the non-parametric Mann-Kendal test and the Mann-Kendal seasonal test, used to determine the significance of trends of the term (Xu et al., 2003). The test relationships used to examine the existence of the trend are as follows:

$$U_{MK} = \begin{cases} (MK - 1) \text{var}(MK)^{-0.5} & MK > 0 \\ 0 & MK = 0 \\ (MK + 1) \text{var}(MK)^{-0.5} & MK < 0 \end{cases} \quad (A4)$$

where U_{MK} is the standard of Mann-Kendall statistic, MK is the Man-Kendall statistic, and $\text{var}(MK)$ is the variance of MK . The MK and $\text{var}(MK)$ are defined as:

$$MK = \sum_{i=1}^{N-1} \sum_{j=i+1}^N \text{sgn}(l_j - l_i) \quad (A5)$$

$$\text{var}(MK) = \left((2N^3 - 7N^2 - 5N) - \sum_j^g O_j(O_j - 1)(2O_j + 5) \right) / 18 \quad (A6)$$

where l is the lake levels, g is the number of identical groups, O_j is the number of observations at the j th group, sgn is the sign function, and N is the number of samples.

- In the second round of preprocessing, the seasonal and non-seasonal differencing and simultaneous seasonal and non-seasonal differencing were employed. The use of concurrent seasonal and non-seasonal differencing not only removed the deterministic terms and stationarized the series, but also lowered the seasonal and non-seasonal differencing to achieve the optimum model.
- The new linear-based method is appropriate for forecasting monthly Lake water levels. The best models for the Van Lake was obtained using standardization preprocessing as the first step and the final model was (1,1,2)(0,1,1)₁₂ ($R^2 = 0.9951$; $RMSE$ (m) = 0.0417; $MAPE\%$ = 0.0019; $RMSRE$ = 0.000025; NS_E = 0.995; AIC_C = -906.36) while for the Michigan-Huron Lake, the best preprocessing was obtained using spectral analysis and the best model had performance metrics of ($R^2 = 0.9889$; $RMSE$ (m) = 0.0368; $MAPE\%$ = 0.0171; $RMSRE$ = 0.00209; NS_E = 0.9887; AIC_C = -1580.11).
- Comparing the proposed linear based methodology for lake level modeling with the ANN and the ANFIS methods using different relative and absolute statistical indices for the accuracy and complexity of model predictions shows that the proposed methodology in the current study outperformed the nonlinear based methods regarding complexity and efficiency. This superior accuracy and lower uncertainty of the new model predictions compared to the ANN and ANFIS models were detected for both the Van Lake and the Michigan-Huron Lake.
- Due to the high performance of the proposed linear based, this approach is recommended for modeling LLTS at other lakes. Moreover, this methodology is suggested for the other time series forecasting applications.

Conflict of interest

The authors declare that they have no conflict of interest.

Also, the seasonal Mann-Kendall tests are expressed as follows:

$$S_k = \sum_{i=1}^{N_k} \sum_{j=i+1}^{N_k-1} \text{sgn}(l_{ki} - l_{kj}) \quad (\text{A7})$$

$$SMK = \sum_{k=1}^{\omega} (S_k - \text{sgn}(S_k)) \quad (\text{A8})$$

$$\text{var}(SMK) = 2 \sum_{i=1}^{\omega-1} \sum_{j=i+1}^{\omega} \sigma_{ij} + \sum_k^{\omega} (2N_k^3 - 7N_k^2 - 5N_k)/18 \quad (\text{A9})$$

$$U_{SMK} = MK \text{var}(MK)^{-0.5} \quad (\text{A10})$$

where ω is the number of the season in a year and σ_{ij} is the covariance of statistic test in season i and j . If the probabilities of these tests are higher than the significant level of 0.05, then time series has no trend.

Appendix C. Jump analysis

The non-parametric Mann-Whitney test is employed to check that the randomly selected values from one sample of time series, are equal or different from the selected ones from the second sample of it and thus the jump in series is assessed (Mann and Whitney, 1947). The statistic of this test is computed as follow:

$$MW_U = \sum_{i=1}^{N_1} (D(O_i) - (m_1(m_1 + m_2 + 1)/2)) / ((m_1 m_2 (m_1 + m_2 + 1))^{0.5}/12) \quad (\text{A11})$$

In this regard, the O_i is series arranged according to the original series $l(t)$, $D(O_i)$, the degree of function O_i , m_1 , and m_2 are the number of sub-series members of the main series that the sum of which is equal to the original series length. If the probability corresponds to the test statistic in the normal probability distribution, the confidence level is greater than α , the assumption of the same distribution of series is acceptable, and the main series is jump free.

Appendix D. Period analysis

The recognition of periodic term as an essential part of the time series is of particular importance which is done with different tests. The Fisher test is used as a method to identify significant periodicities over time. The following equation is used to determine the existence of periodic components (Kashyap and Rao, 1976):

$$F^* = (N(N-2)/4) \frac{\alpha_z^2 + \beta_z^2}{\sum_{z=1}^k (x(t) - \alpha_z \cos(\Omega_z t) - \beta_z \sin(\Omega_z t))} \quad (\text{A12})$$

where F^* , N , α_z , β_z and Ω_z are the Fisher statistic, number of samples, Fourier coefficients and the angular frequency, respectively. The α_z , β_z and Ω_z are defined as follows

$$\alpha_z = \frac{2}{N} \left(\sum_{t=1}^N l(t) \cos(2\pi f_z t) \right) \quad z = 1, 2, \dots, k \quad (\text{A13})$$

$$\beta_z = \frac{2}{N} \left(\sum_{t=1}^N l(t) \sin(2\pi f_z t) \right) \quad z = 1, 2, \dots, k \quad (\text{A14})$$

$$f_z = \frac{z}{N} \quad (\text{A15})$$

$$\Omega_z = \frac{2\pi z}{N} \quad z = 1, 2, \dots, k \quad (\text{A16})$$

where f_z is the z th based frequency harmony.

If the F distribution critical value at a significant level ($F(2, N-2)$) is lower than F^* then The Ω_z related periodicity will not be significant. In other words:

$$F^* \geq F(2, N-2) \quad (\text{A17})$$

Appendix E. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jhydrol.2019.01.010>.

References

- Altunkaynak, A., 2007. Forecasting surface water level fluctuations of Lake Van by artificial neural networks. *Water Resour. Manage.* 21 (2), 399–408. <https://doi.org/10.1007/s11269-006-9022-6>.
- Atieh, M., Taylor, G., Sattar, A.M., Gharabaghi, B., 2017. Prediction of flow duration curves for ungauged basins. *J. hydrol.* 545, 383–394. <https://doi.org/10.1016/j.jhydrol.2016.12.048>.
- Azimi, H., Bonakdari, H., Ebtehaj, I., Khoshbin, F., 2018. Evolutionary design of

- generalized group method of data handling-type neural network for estimating hydraulic jump roller length. *Acta Mech.* 229, 1197–1214. <https://doi.org/10.1007/s00707-017-2043-9>.
- Bonakdari, H., Moeeni, H., Ebtehaj, I., Zeynoddin, M., Mahoammadian, A., Gharabaghi, B., 2018. New insights into soil temperature time series modeling: linear or non-linear? *Theor. Appl. Climatol.* <https://doi.org/10.1007/s00704-018-2436-2>.
- Brinkmann, W., 2000. Causes of variability in monthly Great Lakes water supplies and lake levels. Retrieved from. *Clim. Res.* 15, 151–160. <http://www.jstor.org/stable/24867308>.
- Burnham, K.P., Anderson, D.R., 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, second ed. Springer-Verlag ISBN 0-387-95364-7.
- Çimen, M., Kisi, O., 2009. Comparison of two different data-driven techniques in modeling lake level fluctuations in Turkey. *J. Hydrol.* 378, 253–262. <https://doi.org/10.1016/j.jhydrol.2009.09.029>.
- Ebtehaj, I., Bonakdari, H., Khoshbin, F., Azimi, H., 2015. Pareto genetic design of group method of data handling type neural network for prediction discharge coefficient in rectangular side orifices. *Flow Meas. Instrum.* 41, 67–74. <https://doi.org/10.1016/j.flowmeasinst.2014.10.016>.
- Ebtehaj, I., Bonakdari, H., Moradi, F., Gharabaghi, B., Khozani, Z.S., 2018. An integrated framework of Extreme Learning Machines for predicting scour at pile groups in clear water condition. *Coast. Eng.* 135, 1–15. <https://doi.org/10.1016/j.coastaleng.2017.12.012>.
- Freeman, B.S., Taylor, G., Gharabaghi, B., Thé, J., 2018. Forecasting air quality time series using deep learning. *J. Air Waste Manage. Assoc.* <https://doi.org/10.1080/10962247.2018.1459956>. Accepted author version posted online: 13 Apr 2018.
- Gholami, A., Bonakdari, H., Ebtehaj, I., Shaghghi, S., Khoshbin, F., 2017. Developing an expert group method of data handling system for predicting the geometry of a stable channel with a gravel bed: New model for predicting stable channel geometry with a gravel bed. *Earth Surf. Proc. Land.* 42, 1460–1471. <https://doi.org/10.1002/esp.4104>.
- Güldal, V., Tongal, H., 2010. Comparison of recurrent neural network, adaptive neuro-fuzzy inference system and stochastic models in Eğirdir Lake level forecasting. *Water Resour. Manage.* 24, 105–128. <https://doi.org/10.1007/s11269-009-9439-9>.
- Hirsch, R.M., Slack, J.R., 1984. A nonparametric trend test for seasonal data with serial dependence. *Water Resour. Res.* 20 (6), 727–732. <https://doi.org/10.1029/WR020i006p00727>.
- Hurst, H.E., Black, R.P., Simaika, Y.M., 1965. Long-term storage. An experimental study, Constable, London (1965). BB Mandelbrot.
- Jang, J.S., 1993. ANFIS: adaptive-network-based fuzzy inference system. *IEEE T. Syst. Man. Cyb.* 23, 665–685. <https://doi.org/10.1109/21.256541>.
- Kashyap, R.L., Rao, A.R., 1976. *Dynamic Stochastic Models from Empirical Data* (Vol. illustrated). Mathematics in Science and Engineering. Harcourt Brace Jovanovich (Academic Press), New York and San Francisco, USA, pp. 334.
- Kazemian-Kale-Kale, A., Bonakdari, H., Gholami, A., Khozani, Z.S., Akhtari, A.A., Gharabaghi, B., 2018. Uncertainty analysis of shear stress estimation in circular channels by Tsallis entropy. *Phys. A: Stat. Mech. Appl.* 510, 558–576. <https://doi.org/10.1016/j.physa.2018.07.014>.
- Khozani, Z.S., Bonakdari, H., Ebtehaj, I., 2017. An analysis of shear stress distribution in circular channels with sediment deposition based on Gene Expression Programming. *Int. J. Sedim. Res.* 32, 575–584. <https://doi.org/10.1016/j.ijsrc.2017.04.004>.
- Kiş, Ö., 2009. Neural network and wavelet conjunction model for modelling monthly level fluctuations in Turkey. *Hydrol. Process.* 23, 2081–2092. <https://doi.org/10.1002/hyp.7340>.
- Mann, H.B., Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* 50–60.
- Marco, J.B., Harboe, R., Salas, J.D., 1993. *Stochastic Hydrology and its Use in Water Resources Systems Simulation and Optimization*. Springer Science & Business Media, Dordrecht, Netherlands.
- Moeeni, H., Bonakdari, H., Ebtehaj, I., 2017a. Integrated SARIMA with neuro-fuzzy systems and neural networks for monthly inflow prediction. *Water Resour. Manage.* 31, 2141–2156. <https://doi.org/10.1007/s11269-017-1632-7>.
- Moeeni, H., Bonakdari, H., Ebtehaj, I., 2017b. Monthly reservoir inflow forecasting using a new hybrid SARIMA genetic programming approach. *J. Earth Syst. Sci.* 126, 18. <https://doi.org/10.1007/s12040-017-0798-y>.
- Moradi, F., Bonakdari, H., Kisi, O., Ebtehaj, I., Shiri, J., 2018. Abutment scour depth modeling using neuro-fuzzy embedded techniques. *Mar. Georesour. Geotech.* <https://doi.org/10.1080/1064119X.2017.1420113>.
- Myronidis, D., Ioannou, K., Fotakis, D., Dörfinger, G., 2018. Streamflow and hydrological drought trend analysis and forecasting in cyprus. *Water Resour. Manage.* 2 (5), 1759–1776. <https://doi.org/10.1007/s11269-018-1902-z>.
- Myronidis, D., Stathis, D., Ioannou, K., Fotakis, D., 2012. An integration of statistics temporal methods to track the effect of drought in a shallow Mediterranean Lake. *Water Resour. Manage.* 26 (15), 4587–4605. <https://doi.org/10.1007/s11269-012-0169-z>.
- Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. *J. Hydrol.* 10, 282–290. [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6).
- Sanikhani, H., Kisi, O., Kiafar, H., Ghavidel, S.Z.Z., 2015. Comparison of different data-driven approaches for modeling lake level fluctuations: the case of Manyas and Tuz Lakes (Turkey). *Water Resour. Manage.* 29, 1557–1574. <https://doi.org/10.1007/s11269-014-0894-6>.
- Shabanlou, S., 2018. Improvement of extreme learning machine using self-adaptive evolutionary algorithm for estimating discharge capacity of sharp-crested weirs located on the end of circular channels. *Flow Meas. Instrum.* 59, 63–71. <https://doi.org/10.1016/j.flowmeasinst.2017.11.003>.
- Shafaei, M., Kisi, O., 2016. Lake level forecasting using wavelet-SVR, wavelet-ANFIS and wavelet-ARMA conjunction models. *Water Resour. Manage.* 30, 79–97. <https://doi.org/10.1007/s11269-015-1147-z>.
- Shaghghi, S., Bonakdari, H., Gholami, A., Ebtehaj, I., Zeinolabedini, M., 2017. Comparative analysis of GMDH neural network based on genetic algorithm and particle swarm optimization in stable channel design. *Appl. Math. Comput.* 313, 271–286. <https://doi.org/10.1016/j.amc.2017.06.012>.
- Shiri, J., Shamshirband, S., Kisi, O., Karimi, S., Bateni, S.M., Nezhad, S.H.H., Hashemi, A., 2016. Prediction of water-level in the Urmia Lake using the extreme learning machine approach. *Water Resour. Manage.* 30, 5217–5229. <https://doi.org/10.1007/s11269-016-1480-x>.
- Siegel, S., Castellan, N.J., 1988. *The Case of k Related Samples. Nonparametric Statistics for Behavioral Sciences*. McGraw-Hill, New York, pp. 170–174.
- Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.-Atmos.* 106, 7183–7192. <https://doi.org/10.1029/2000JD900719>.
- Traore, S., Luo, Y., Fipps, G., 2017. Gene-Expression Programming for Short-Term Forecasting of Daily Reference Evapotranspiration Using Public Weather Forecast Information. *Water Resour. Manage.* 31, 4891–4908. <https://doi.org/10.1007/s11269-017-1784-5>.
- UNESCO, 2003. http://www.unesco.org/water/wwap/wwdr/wwdr1/table_contents/index.html.
- Valipour, M., Banihabib, M.E., Behbahani, S.M.R., 2013. Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. *J. Hydrol.* 476, 433–441. <https://doi.org/10.1016/j.jhydrol.2012.11.017>.
- Wilcox, D.A., Thompson, T.A., Booth, R.K., Nicholas, J.R., 2007. Lake-level variability and water availability in the Great Lakes. U.S. Geological Survey Circular 1311. *Environ. Sci. Biol. Faculty Publ.* 25. https://pubs.usgs.gov/circ/2007/1311/pdf/circ1311_web.pdf.
- Wang, W.C., Chau, K.W., Cheng, C.T., Qiu, L., 2009. A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. *J. Hydrol.* 374, 294–306. <https://doi.org/10.1016/j.jhydrol.2009.06.019>.
- Wang, W.C., Chau, K.W., Xu, D.M., Chen, X.Y., 2015a. Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition. *Water Resour. Manage.* 29, 2655–2675. <https://doi.org/10.1007/s11269-015-0962-6>.
- Wang, W.C., Chau, K.W., Qiu, L., Chen, Y.B., 2015b. Improving forecasting accuracy of medium and long-term runoff using artificial neural network based on EEMD decomposition. *Environ. Res.* 139, 46–54. <https://doi.org/10.1016/j.envres.2015.02.002>.
- Xu, Z.X., Takeuchi, K., Ishidaira, H., 2003. Monotonic trend and step changes in Japanese precipitation. *J. Hydrol.* 279, 144–150. [https://doi.org/10.1016/S0022-1694\(03\)00178-1](https://doi.org/10.1016/S0022-1694(03)00178-1).
- Zaji, A.H., Bonakdari, H., Gharabaghi, B., 2018. Reservoir water level forecasting using group method of data handling. *Acta Geophysica.* 66 (4), 717–730. <https://doi.org/10.1007/s11600-018-0168-4>.
- Zeynoddin, M., Bonakdari, H., Azari, A., Ebtehaj, I., Gharabaghi, B., Madavar, H.R., 2018. Novel hybrid linear stochastic with non-linear extreme learning machine methods for forecasting monthly rainfall a tropical climate. *J. Environ. Manage.* 222, 190–206. <https://doi.org/10.1016/j.jenvman.2018.05.072>.